Extensión de métodos modernos de Aprendizaje Automatizado y aplicaciones.

Hernán C. Ahumada, Ariel E. Bayá, Guillermo L. Grinblat, C. Javier Izetta, Lucas C. Uzal y Pablo M. Granitto,
CIFASIS (CONICET – UN Rosario - UPCAM)
Bv 27 de Febrero 210 bis – 2000 Rosario
(341) 4237248 int 303
{ahumada,baya,grinblat,izetta,uzal,granitto}@cifasis-conicet.gov.ar

Resumen

El campo del Aprendizaje Automatizado (Machine Learning) es parte central de la nueva revolución tecnológica basada en el uso inteligente de la información. Por tradición, los principales problemas que se investigan en esta área son los de reconocimiento de patrones o Clasificación, aproximación de funciones de variable continua o Regresión, y búsqueda de estructuras ocultas en datos o Clustering. Lógicamente, el desarrollo de nuevos métodos y algoritmos se concentró en un principio en los problemas más simples o típicos de encontrar, por ejemplo en problemas estacionarios en el tiempo, con una abundante cantidad de ejemplos de los cuales aprender y con sólo unas pocas clases bastante balanceadas entre sí. Sin embargo, los nuevos tipos de datos provenientes de la genómica, la proteómica, los equipos de monitoreo continuo de sistemas críticos, etc., han introducido nuevos desafíos en el Aprendizaje Automatizado. Este proyecto propone el desarrollo de nuevos métodos (o la extensión de los métodos actuales cuando sea apropiado) para poder modelar eficientemente esta nueva clase de datos, incluyendo problemas de regresión y clasificación no estacionarios y/o con gran nivel de ruido, problemas de clasificación y clustering con un número extremadamente alto de variables de entrada, o problemas de clasificación con un importante desbalance entre clases. En todas las líneas del proyecto se incluyen aplicaciones a problemas actuales de gran interés tecnológico, como la biotecnología y la agrotecnología.

Palabras clave: Machine Learning, Clasificación, Regresión, Clustering

Contexto

Todas las líneas de investigación forman parte de un único proyecto financiodo por ANPCyT (PICT 237/2008). El proyecto se desarrolla en el grupo de Machine Learning y Aplicaciones del CIFASIS

Introducción

El campo del Aprendizaje Automatizado (“Machine Learning”) es parte central de la

nueva revolución tecnológica basada en el uso inteligente de la información. Su objetivo primario es la construcción de algoritmos que automáticamente mejoren su eficiencia en la solución de un problema a través de la experiencia acumulada (Mitchell97). El Aprendizaje Automatizado integra conocimientos de disciplinas diversas como Inteligencia Artificial, Estadística, Ciencia Cognitiva, Neurobiología, etc.

Tradicionalmente, los principales problemas que se investigan en esta área (Duda00) son los de reconocimiento de patrones o Clasificación, aproximación de funciones de variable continua o Regresión, y descubrimiento de estructuras ocultas en datos o Clustering. Lógicamente, el desarrollo de nuevos métodos y algoritmos se concentró en los problemas más simples o típicos de encontrar, por ejemplo en problemas estacionarios en el tiempo, con una abundante cantidad de ejemplos de los cuales aprender y con sólo unas pocas clases bastante balanceadas entre sí. Sin embargo, los nuevos tipos de datos provenientes de la genómica, la proteómica, los equipos de monitoreo continuo de sistemas críticos, etc., han introducido nuevos desafíos en el Aprendizaje Automatizado. Este proyecto propone como objetivo general el desarrollo de nuevos métodos (o la extensión de los métodos actuales cuando sea apropiado) para poder abordar eficientemente esta nueva clase de datos, incluyendo problemas de regresión y clasificación no estacionarios y/o con gran nivel de ruido, problemas de clasificación y clustering con un número extremadamente alto de variables de entrada, o problemas de clasificación con un importante desbalance entre clases. Todas las aplicaciones tienen relación con problemas tecnológicos actuales de gran importancia en áreas como la biotecnología, la agrotecnología y la industria siderúrgica.

En Aprendizaje Automatizado se conoce como problema de clases desbalanceadas a aquel problema de clasificación en que al menos una de las clases es claramente minoritaria o está sub-representada respecto de las otras. Esta característica es frecuente en problemas de gran relevancia como el diagnóstico de enfermedades, la identificación de señales o la predicción de fallas (Fawcett97, Chawla02).

La detección de novedades, o clasificación de una clase (Shawe-Taylor04), ha cobrado interés recientemente. En estos problemas se trata de detectar comportamientos “anormales” de un sistema observando solamente los comportamientos “normales” del mismo. En cierto sentido se puede considerar a esta clase de problemas como el límite de los problemas desbalanceados, donde una clase es tan rara que la aparición de un ejemplo es considerado como una “novedad”.

Para los sistemas llamados complejos el conjunto de variables relevantes es en general desconocido. Sin embargo, si se estudia un sistema determinista efectuando mediciones sobre el mismo a intervalos regulares de tiempo, es decir, a través de series temporales, los valores pasados de dicha serie sirven como substituto de las verdaderas variables de estado del sistema en estudio. Un resultado de Takens (Takens81) muestra que, si la cantidad de datos disponibles es infinita y están libres de ruido observacional, entonces para casi todo tiempo de retraso τ y dimensión de "embedding" d suficientemente grande, el espacio $[x(t), x(t-\tau), x(t-2\tau),... x(t-(d-1)\tau)]$ es equivalente al espacio de estados del sistema dinámico original. En la situación real de una cantidad finita de registros afectados de cierta incertidumbre, no todos los

valores posibles de τ y d son igualmente adecuados. La determinación de valores óptimos para estas dos magnitudes es objeto de estudio en la literatura.

La selección de características, variables o “feature selection”, es una técnica de pre-procesado que se aplica normalmente en aprendizaje automatizado a conjuntos de datos de alta dimensionalidad. Esta técnica estudia cómo seleccionar eficientemente un subconjunto de atributos o variables que se usarán luego para construir los modelos. La selección de variables tiene diversos propósitos: reducir la dimensión del espacio de variables (“curse of dimensionality”), eliminar variables irrelevantes o redundantes y mejorar la eficiencia de los algoritmos de aprendizaje, o mejorar la interpretabilidad de los modelos, entre otros (Guyon03).

Líneas de investigación y desarrollo

I. Clasificación con clases altamente desbalanceadas.

En esta línea se propone el desarrollo de métodos mixtos que incluyen una primera etapa en la que se descompone el problema en sub-problemas más simples, en forma no supervisada, y una segunda etapa en que se desarrolla un conjunto de clasificadores, uno para cada subproblema (Ahumada08a). Se propone también la aplicación de los nuevos desarrollos a problemas reales de importancia, como estimación de riesgo quirúrgico o la detección de proteínas homólogas.

II. Detección de novedades en problemas no estacionarios.

En esta segunda línea de trabajo se desarrollan métodos de clasificación de una clase para problemas no-estacionarios con variación lenta en el tiempo, lo cual es una continuación de los trabajos previos en clasificadores no estacionarios (Grinblat08a). Esta línea está conectada a una posible aplicación real en predicción de fallas en equipos críticos, en colaboración con la empresa Ternium-Siderar.

III Regresión en sistemas complejos

Proponemos en este proyecto desarrollar e implementar técnicas de "embedding" no uniformes (en los que ciertos valores retrasados de x están ausentes) para el análisis y predicción de series temporales de naturaleza caótica, comparándolas con las existentes en la literatura para el enfoque uniforme (que incluye todas las variables retrasadas hasta la dimensión de embedding). Se planea aplicar las metodologías desarrolladas al estudio del fenómeno de El Niño (índice de oscilación sur), comportamiento de la capa de ozono, y temperatura media de la Tierra, entre otras series temporales asociadas a sistemas reales.

IV Selección de características.

La última línea de investigación de este proyecto es extender adecuadamente el método RFE (Guyon02) a problemas de clasificación no-estacionarios en el tiempo. Un segundo objetivo es este área es desarrollar métodos estables de selección para problemas con muchas variables correlacionadas (como por ejemplo los datos de expresión génica o de espectrometría de masa), mediante el uso de una primera etapa de clustering de variables que utilice métodos específicamente desarrollados para esa tarea.

Formación de Recursos Humanos

El equipo de trabajo está formado actualmente por un investigador formado (PMG), dos post-docs con becas de CONICET (AEB y GLG) y tres doctorandos (HCA, CJI y LCU) becados por CONICET. Hay 5 tesis doctorales que forman parte del proyecto, dos fueron defendidas en Marzo 2011 y otras 3 están actualmente en desarrollo.

Referencias

(Ahumada 08a) H. Ahumada, G.L. Grinblat, L.C. Uzal, P. M. Granitto & H.A. Ceccatto, "REPMAC: A new hybrid approach to higly imbalanced classification problems", Proceedings of the Eighth International Conference on Hybrid Intelligent Systems HIS08, Barcelona, Spain, 2008.

(Chawla02) N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, "Smote: Synthetic minority over-sampling technique", Journal of Artificial Intelligence Research, 16, 351–357, 2002.

(Duda00) R. O. Duda, P. E. Hart & D. G. Stork. Pattern Classification, Second Edition. John Wiley & Sons, 2000.

(Fawcett97) R. E. Fawcett & F. Provost, "Adaptive fraud detection", Data Mining and Knowledge Discovery, 3(1), 291–316, 1997.

(Grinblat08a) G. L. Grinblat, P. M. Granitto, A. Ceccatto, "Time-Adaptive Support Vector Machines", Revista Iberoamericana de Inteligencia Artificial, 40, 39, 2008.

(Guyon02) I. Guyon, J. Weston, S. Barnhill & V. Vapnik, "Gene selection for cancer classification using support vector machines", Machine Learning, 46:1-3, 389-422, 2002.

(Guyon03) I. Guyon & A. Elisseeff, "An Introduction to Variable and Feature Selection", Journal of Machine Learning Research, 3(Mar), 1157--1182, 2003.

(Mitchell97) T. Mitchell, Machine Learning, McGraw-Hill, 1997.

(Shawe-Taylor04) J. Shawe-Taylor & N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

(Takens81) F. Takens, "Detecting strange attractors in turbulence", in Dynamical Systems and Turbulence, Lecture Notes in Mathematics 898, 366, Springer-Verlag, 1981.